[56] Page GP, George V, Go RC *et al*. ''Are we there yet?'': Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am J Hum Genet* 2003;**73**:711–19.

[57] Ioannidis JP. Commentary: grading the credibility of molecular evidence for complex diseases. *Int J Epidemiol* 2006; [Epub ahead of print] PMID: 16540537.

[58] Genetic Association Information Network (GAIN). Accessed online on 10/4/2006 at: http://www.fnih.org/GAIN/GAIN_home.shtml.

[59] The Wellcome Trust Case Control Consortium. (WTCCC). Accessed online on 10/4/2006 at: http://www.wtccc.org.uk/.

[60] Rebbeck TR, Spitz M, Wu X. Assessing the function of genetic variants in candidate gene association studies. *Nat Rev Genet* 2004;**5**:589–94.

[61] Jais PH. How frequent is altered gene expression among susceptibility genes to human complex disorders? *Genet Med* 2005;**7**:83–96.

[62] Campbell H, Rudan I. Interpretation of genetic association studies in complex disease. *Pharmacogen J* 2002;**2**:349–60.

[63] Khoury MJ, Newill CA, Chase GC. Epidemiologic evaluation of screening for risk factors: application to genetic screening. *Am J Public Health* 1985;**75**:1204–08.

[64] Pepe MS, Janes H, Longton G *et al*. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004;**159**:882–90.

[65] Holtzman NA, Marteau T. Will genetics revolutionize medicine? *New Engl J Med* 2000;**343**:141–44.

[66] Davey-Smith G, Ibrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32**:1–22.

[67] Khoury MJ, Davis RL, Gwinn M *et al*. Do we need genomic research for the prevention of common diseases with environmental causes? *Am J Epidemiol* 2005;**161**:799–805.

[68] Nitsch D, Molokhia M, Smeeth L *et al*. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *Am J Epidemiol* 2006;**163**:397–403.

[69] Khoury MJ, Yang Q, Gwinn M *et al*. An epidemiologic assessment of genomic profiling for measuring susceptibility to common diseases and targeting interventions. *Genet Med* 2004;**6**:38–47.

[70] Public Population Project in Genomics (P3G). Accessed online on October 4, 2006 at: http://www.p3gconsortium.org/index.cfm.

Published by Oxford University Press on behalf of the International Epidemiological Association

International Journal of Epidemiology 2007;**36**:445–448
doi:10.1093/ije/dym055

Commentary: Rare alleles, modest genetic effects and the need for collaboration

H Campbell* and T Manolio[4]

Accepted 1 March 2007

The article by Khoury *et al*.[1] presents a useful overview of some of the complex issues facing those trying to identify genetic variants underlying common complex disease. They focus on the common disease—common variant model where effect sizes associated with individual genetic variants are small. Undoubtedly this will be the case for most, but not all, variants. An L-shaped or exponential distribution of mutation effect sizes has wide support [2–4] with many variants with small effects, a smaller number with intermediate effects and relatively few with large effects. It could be argued that the genetic variants related to human disease that have been identified to date primarily reflect the study designs used to identify them. Linkage studies conducted among families with multiple cases of disease were successful in identifying highly penetrant variants with large effects. Association studies conducted in general population samples using common genetic markers typically find low penetrance variants with (very) small effects, as noted by Khoury. This is not unexpected given that these common genetic variants are ancient and will have been subject to some selective pressure over time.[3]

We can predict that re-sequencing studies in the near future which study rarer variants (say 0.05–5%) will identify many variants of intermediate effect associated with common complex disease. This paradigm shift has already begun with the seminal work of Cohen, who compared non-synonymous sequence variations in individuals at the extremes of the population distribution of LDL-cholesterol levels, and determined that a significant fraction of genetic variance is due to multiple alleles with intermediate effects that are present at low frequencies (0.05–5%) in the population, particularly persons of African ancestry.[5] Until many such studies are reported it will be premature to decide on the relative importance of the common variant—common disease model and the alternative rare variant—common disease model which states that disease susceptibility to common diseases is the result of multiple low frequency/rare variants with larger phenotypic effects. As Cohen notes, although individually rare, these variants may be

* Corresponding author. Division of Community Health Sciences, Public Health Sciences, University of Edinburgh, Teviot Place, Edinburgh EH8 9AG, UK. E-mail: harry.campbell@ed.ac.uk

collectively common in the population. This has important consequences since the issues of causal inference and clinical application, described well by Khoury, maybe somewhat different for these variants. Our view of the genetic contribution to common complex disease is becoming clearer but is still biased as it is highly determined by available genetic technology and its cost.[6]

Nevertheless, the questions raised by Khoury with respect to the common variants currently employed in genetic association studies are important to consider. The most immediate questions are perhaps how to interpret the findings of the increasing number of genome-wide association (GWA) studies which are now being conducted and how this global research effort could be most efficiently marshalled.

Interpretation of findings from GWA studies

There is currently a need to share experience in the design and analysis of GWA studies. This needs to address issues such as whether common controls are workable; what sample size in phase 1 of two stage studies is required to give adequate power to separate false from true positives; what is the best way to rank results to select variants to take to subsequent phases; how to combine data across studies and then integrate this with other information. Efforts to define best practice in design and analysis are underway and would benefit from the support of groups such as Human Genome Epidemiology Network (HuGENet) which can use their convening power to bring international groups together to tackle these issues.

It is of concern that few studies have followed initial reports of disease associations with the identification of the causal genetic variant. We need an approach to ranking reported associations in terms of their likelihood of being causal so that this can be used to prioritize future research investment. The Bradford Hill criteria still provide a useful framework for considering causal inference.[7] Biological plausibility, through bioinformatics interrogation of biological databases to assess impact on amino acid sequence and subsequent protein structure and function or to investigate the degree of genetic conservation across species,[8] can sometimes provide data strongly against a causal role but rarely gives compelling support in favour of causality. Until a few years ago, it was generally considered that experimental data for example from animal models or gene expression studies would yield clear causal information. However, this has recently been challenged and recommendations given that there is a need for caution and for a priori hypotheses when citing biological or functional data as supportive evidence.[9] Strength of association as demonstrated by a genetic variant showing a large effect size would remain important evidence but, as Khoury notes, the typical effect sizes for common variants in complex disease have been in the order of 1.1–1.5, at the limits of resolution of epidemiological studies. New approaches such as Mendelian randomization[10,11] and integrative genomics (the joint assessment of gene function and expression)[12] hold promise of providing more robust information on causality but at present replication has become the criterion that has assumed most importance.

It is clear that a major challenge in GWA studies is the extremely low prior probability for a given single nucleotide polymorphism (SNP) (among hundreds of thousands tested) to have a causal role in the disease under study. This means that individual studies will have low power to distinguish between true and false positives. The need for replication of findings becomes paramount and it quickly becomes apparent that networks of investigators need to tackle the problem together. False negative results are an important problem with replication studies.[13] This is, in part, due to power being overestimated based on upwardly biased effect sizes (due to the 'winners curse' phenomenon) and failure to account for genetic heterogeneity (different causal genes in different individuals) and aetiological heterogeneity. It is clear that, for replication, there is a need for very large case-control collections with $>10\,000$ cases and controls across a collaborative network of studies rather than many small underpowered studies where only a biased sample are published.[14]

Need for planned international collaboration and data sharing

Khoury *et al*. make a strong case for the development of standards for presenting and interpreting gene–disease associations, and have made similar calls in the past.[15,16] Chief among their reasons for doing so is to permit valid and robust syntheses of available evidence, preferably through true meta-analyses, to remedy the many shortcomings of the existing literature. These shortcomings include preferential publication of positive findings, underpowered and potentially biased study samples that increase the likelihood of erroneous reports and the tendency to declare definitive associations on the basis of a single study.[17] Standardized and complete reporting of these studies, though potentially cumbersome, would permit more reliable and objective assessment of the evidence for or against a proposed association, and even the possibility of grading or quality-scoring individual reports.[16] More importantly, standardization would permit ready syntheses of the published literature, particularly for differential associations among subgroups or for gene–environment interactions in which even the largest study is likely to have limited power.

As valuable as standardized reporting and meta-analyses can be, they probably cannot take the place of meaningful communication and interaction among investigative groups. Khoury and colleagues[18] have made a strong case in the past, and continue to do so now, for collaboration among investigators to pool and compare gene–disease data, recognizing that even large association studies are likely to be underpowered for genes of modest effect. Collaboration among investigators through disease-related networks, or even across diseases in the proposed 'Network of Networks' approach promoted by the HuGENet, holds the potential for speeding the replication of true associations and rapidly setting aside those that are spurious.[19]

Rapid, unrestricted access to gene-disease association data is becoming an expectation of GWA studies, building on the

strong foundation laid by the Human Genome Project in the Ft. Lauderdale agreement (http://www.wellcome.ac.uk/doc_wtd003208.html). Leading the way in this rapid data access model has been the U.S. National Cancer Institute's 'Cancer Genetic Markers of Susceptibility (CGEMS)' project (http://cgems.cancer.gov/data/), which provided detailed, multivariate adjusted and unadjusted association statistics on over 300 000 SNP markers with prostate cancer in October 2006, as soon as the data were cleaned and released to the participating investigators for analysis. This group released an additional 240 000 SNPs in the same pre-computed format, again immediately after data cleaning, in February 2007. The National Institute of Neurologic Diseases and Stroke provided a similar model in releasing data from its Parkinson's disease genome-wide scan immediately upon publication,[20] and the National Heart, Lung and Blood Institute's Framingham Heart Study has announced it will release association data one year after completion of genome-wide genotyping in over 9000 participants in three generations (http://www.nhlbi.nih.gov/new/press/06-02-06.htm).[21] Two upcoming programmes led by the National Human Genome Research Institute, the Genetic Association Information Network (GAIN, http://www.fnih.org/GAIN/GAIN_home.shtml) and the Genes and Environment Initiative (GEI, http://genesandenvironment.nih.gov/) will release grouped genotype–phenotype association findings publicly as soon as genotyping is completed, and will provide de-identified individual genotype and phenotype data to qualified researchers agreeing to protect participant confidentiality and to abide by other study policies on publication and intellectual property. A similar plan for widespread data release is being implemented in the Wellcome Trust Case Control Consortium, which began releasing de-identified individual genotype data on its control subjects in November, and will release genotype data on its multiple case groups within the next 6 months (http://www.wtccc.org.uk/info/access_to_data_samples.shtml).

This surge in data distribution among individual studies and Institutes has led the National Institutes of Health (NIH) as a whole to develop policies for data release in GWA studies, subject to appropriate human subjects protections. Following a lengthy public commentary and consultation process (http://grants.nih.gov/grants/gwas/background.htm), the NIH has nearly completed its GWA data sharing policies and expects to announce them this spring. To receive these data and provide mechanisms for rapid access, the National Library of Medicine, one of the NIH's 27 Institutes and Centres, has developed the Database of Genotypes and Phenotypes (dbGaP) modelled on the public repository 'dbSNP' for single nucleotide polymorphism data (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gap). This new database will go a long way toward meeting the call of Khoury and colleagues[22] for deposition of genotype–phenotype data in standard formats, and will provide much more in terms of study protocols, forms and other documentation. There may also be a need to produce meta-analyses of data in these databases in the form of regularly updated cumulative odds ratio estimates. This shares some similarities to the role of scientific curators established for databases related to specific Mendelian mutations that result in human disease. Recognition of the importance and funding for the support of this function will be important.

Rapid, widespread availability of GWA data, in addition to facilitating collaborations and reducing the impact of publication bias, will also facilitate rapid replication of findings. Since replication has been called the *sine qua non* of genetic association studies,[9] and recognizing the small number of candidate gene studies that have been replicated,[23] rapid evaluation of a GWA finding in another, similar data set would be tremendously useful in evaluating the importance of a putative association. In the past, such replication efforts have often had to await the development of collaborations and the exchange of samples or reagents, as anticipated in the proposed 'roadmap for reliable human genome studies'.[15] With the advent of databases such as dbGaP, however, adequately documented and reported studies could conceivably be evaluated for replication very rapidly indeed. Critical to the valid interpretation of such comparisons will be an understanding of the potential biases and unique characteristics of each sample set, in terms of differences in selection criteria, case definition, treatment, ancestry, environmental exposures, etc. all of which may profoundly affect the GWAs detected.[24] Equally important will be an understanding of the origin and potential biases of the controls, recognizing that a valid control group should arise from a similar genetic and environmental background as cases, be representative of persons at risk for the disease, and have the same likelihood of being detected as a case (were they to develop the disease) as did the cases included in the study.[24]

While many of the requests for genotype and phenotype data on individual participants are expected to come from investigators deeply involved in identifying genetic variants related to human disease, such as those who would participate in the HuGENet Network of Networks,[15,17] others may well be interested only in the association data, to compare with their own preliminary findings or to animal or functional studies to determine the potential importance of their own results. Although such uses may not be captured directly by formal citations or collaborations, they are likely to be critical in accelerating the progress of gene–disease research and in allowing related fields to identify in the most productive future directions. For these reasons, the rapid development and expansion of standardized databases for reporting GWA findings is a welcome advance that can be expected to promote scientific rigour in a speedy and cost-efficient manner.

Genetic epidemiology is now entering a phase in which progress will be determined not only by the availability of suitable and affordable genetic technology and appropriate statistical tools, but also the extent to which research groups pool resources and expertise. Investigators' willingness to share and collaborate in this way will in turn require, among other things, a system for recognizing and rewarding all research partners. Khoury *et al*. have proposed some outstanding approaches for facilitating this work, which should be embraced and implemented to move this field forward.

References

[1] Khoury MJ, Little J, Gwinn, Ioannidis JP. On the synthesis and Interpretation of consistent but weak gene–disease association in the era of genome-wide association studies. *Int J Epidemiol* 2007;**36**:439–45.

[2] Barton NH, Keightley PD. Understanding quantitative genetic variation. *Nat Rev Genet* 2002;**3**:11–21.

[3] Hayes B, Goddard ME. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet Select Evol* 2002;**33**:209–30.

[4] Wright A, Charlesworth B, Rudan I, Carothers A, Campbell H. A polygenic basis for late-onset disease. *Trends Genet* 2003;**19**:97–106.

[5] Cohen JC, Pertsemlidis A, Fahmi S, *et al*. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. *Proc Natl Acad Sci USA* 2006;**103**:1810–15.

[6] Collins FS. 2005 William Allan Award address. No longer just looking under the lamppost. *Am J Hum Genet* 2006;**79**:421–26.

[7] Campbell H, Rudan I. Interpretation of genetic association studies in complex disease. *Pharmacogenomics J* 2002;**2**:349–60.

[8] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet* 2003;**33**:228–37.

[9] Todd JA. Statistical false positive or true disease pathway? *Nat Genet* 2006;**38**:731–33.

[10] Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003;**32**:1–22.

[11] Nitsch D, Molokhia M, Smeeth L, DeStavola BL, Whittaker JC, Leon DA. Limits to causal inference based on Mendelian randomization: a comparison with randomized controlled trials. *Am J Epidemiol* 2006;**163**:397–403.

[12] Schadt EE, Lamb J, Yang X, *et al*. An integrative genomics approach to infer causal associations between gene expression and disease. *Nat Genet* 2005;**37**:710–17.

[13] Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to common disease. *Nat Genet* 2003;**33**:177–82.

[14] Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. *Lancet* 2003;**361**:865–72.

[15] Ioannidis JP, Gwinn M, Little J, *et al*. Human Genome Epidemiology Network and the Network of Investigator Networks. A road map for efficient and reliable human genome epidemiology. *Nat Genet* 2006;**38**:3–5.

[16] Little J, Bradley L, Bray MS, *et al*. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. *Am J Epidemiol* 2002;**156**:300–10.

[17] Ioannidis JP. Why most published research findings are false. *PLoS Med* 2005;**2**:e124.

[18] Ioannidis JP, Bernstein J, Boffetta P, *et al*. A network of investigator networks in human genome epidemiology. *Am J Epidemiol* 2005;**162**:302–4.

[19] Seminara D, Khoury MJ, O'Brien TR, *et al*. Human Genome Epidemiology Network; the Network of Investigator Networks. The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology* 2007;**18**:1–8.

[20] Fung HC, Scholz S, Matarin M, *et al*. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol* 2006;**5**:911–16.

[21] Herbert A, Lenburg ME, Ulrich D, Gerry NP, Schlauch K, Christman MF. Open-access database of candidate associations from a genome-wide SNP scan of the Framingham Heart Study. *Nat Genet* 2007;**39**:135–36.

[22] Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001;**29**:306–9.

[23] Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002;**4**:45–61.

[24] Manolio TA, Bailey-Wilson JE, Collins FS. Genes, environment, and the value of prospective cohort studies. *Nat Rev Genet* 2006;**7**:812–20.